Leveraging History for Faster Sampling of Online Social Networks

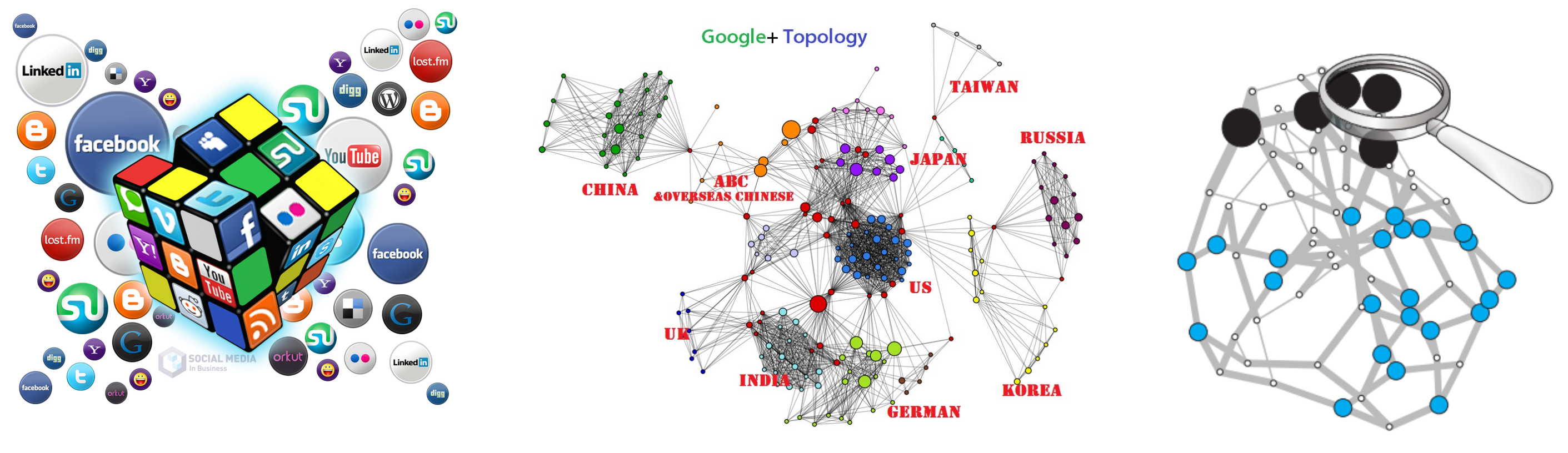
Zhuojie Zhou[1], Nan Zhang[1]

[1]Computer Science Department, George Washington University

School of Engineering & Applied Science
THE GEORGE WASHINGTON UNIVERSITY

Introduction

- **Motivation:** To enable third-party analytical applications of Online Social Networks (OSNs), one must be able to accurately estimate big-picture aggregates (e.g., the AVERAGE age of users, the COUNT of user posts that contain a given word) by issuing a small number of individual-user queries through the social network's web interface.



(a) OSNs examples (b) The graph topology of OSNs (c) Random walk sampling

Figure : Random walk based sampling on OSNs

- **Problem Definition:** How to sample nodes from large graphs using random walks via graph browsing interface with limited query cost while obtain as accurate estimation as possible?
- **Our ideas: History-Aware Random Walks**. The focus of this paper is to offer a *"drop-in" replacement* for this core design (of random walk), such that existing sampling-based analytics techniques over online social networks, no matter which analytics tasks they support or graph topologies they target, can have a better efficiency by leverage random walks' history.

Preliminaries

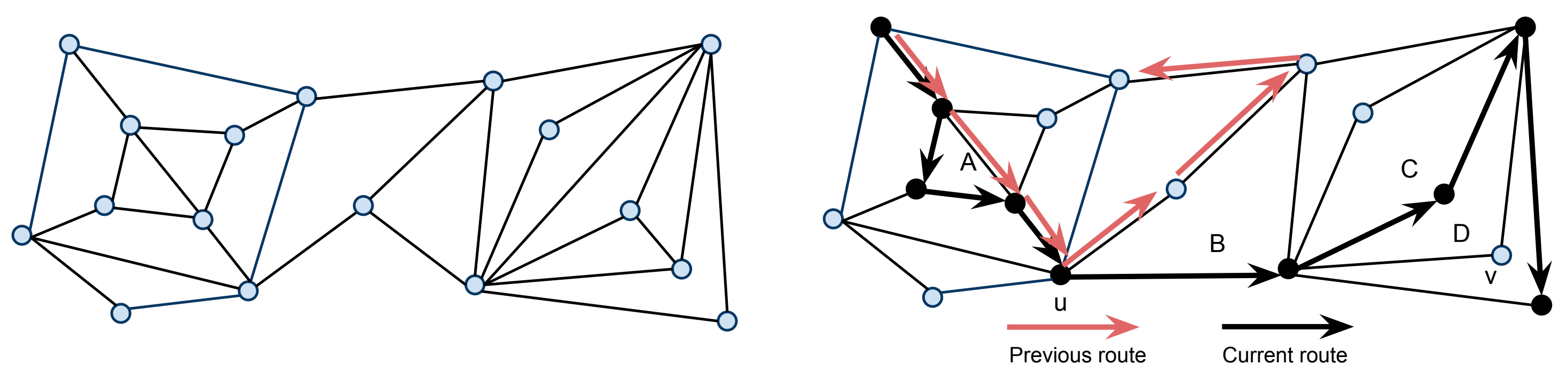
- **Graph Browsing Interface.** The only access channels we have over the data is the web and/or API interface provided by OSNs. While the design of such interfaces varies across different real-world online social networks, almost all of them support queries that take any user ID u as input and return two types of information about u:
  - $N(u)$, the set of all neighbors of u, and
  - all other attributes of u (e.g., user self-description, profile, posts).
- **Random Walk. [Simple Random Walk (SRW)].** Given graph $G(V, E)$, and a node $v \in V$, a random walk is called Simple Random Walk if it chooses uniformly at random a neighboring node $u \in N(v)$ and transit to u in the next step.

$$P_{vu} = \begin{cases} 1/k_v & \text{if } u \in N(v), \\ 0 & \text{otherwise.} \end{cases}$$

That is, SRW selects each node in the graph with probability proportional to its degree.

- **History-Aware Random Walks (Higher order Markov Chain).**

$$\begin{aligned} &\Pr(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) \\ =&\Pr(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_{n-m} = x_{n-m}) \end{aligned}$$



(a) Graph G (b) History-aware random walks will pick different routes whenever possible to avoid stuck in the same region of the graph.

Figure : A demo shows a History-aware random walks on-the-fly.

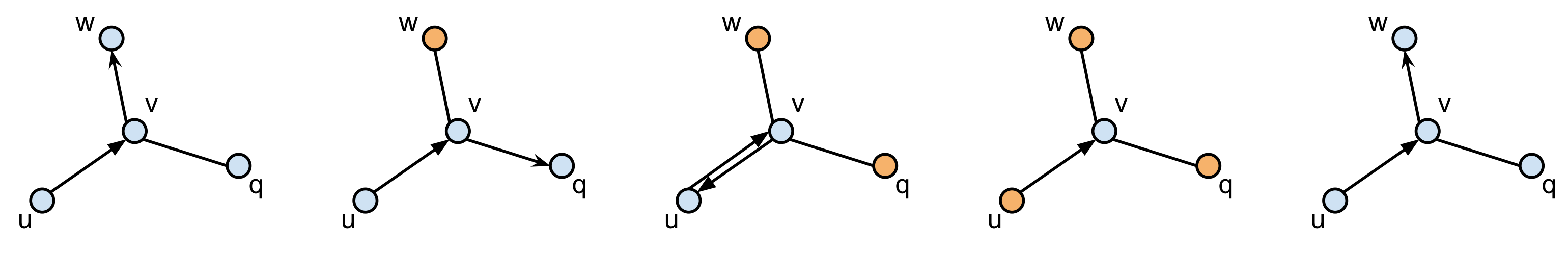
CNRW: Circulated Neighbors Random Walk

- **Key idea.** The key idea of CNRW is to replace such a memoryless transition of SRW to a stateful process. Specifically, given the previous transition of the random walk $u \to v$, instead of selecting the next node to visit by *sampling with replacement* from $N(v)$, i.e., the neighbors of v, we perform such sampling by circulating all v's neighbors *without replacement*.
- **Theorems**
  - Theorem 1. CNRW has the same stationary distribution $\pi(v) = k_v/2|E|$ as SRW's.
  - Theorem 2. The asymptotic variance of CNRW is no greater than SRW's.

$$V_\infty(\hat{\mu}') \leq V_\infty(\hat{\mu}).$$

  - Theorem 3. For a barbell graph, the transition probability

$$\frac{P_{CNRW}}{P_{SRW}} > \frac{|G_1|}{|G_1| - 1} \ln |G_1|.$$



(a) $N(v)$ (b) $N(v) - \{w\}$ (c) $N(v) - \{w, q\}$ (d) $N(v) - \{w, q, u\}$ (e) $N(v)$

Figure : Demo of CNRW, it chooses the next candidate from the set $N(v)$ in a round-robin manner, e.g. circulating the nodes $\{w, q, u\}$.

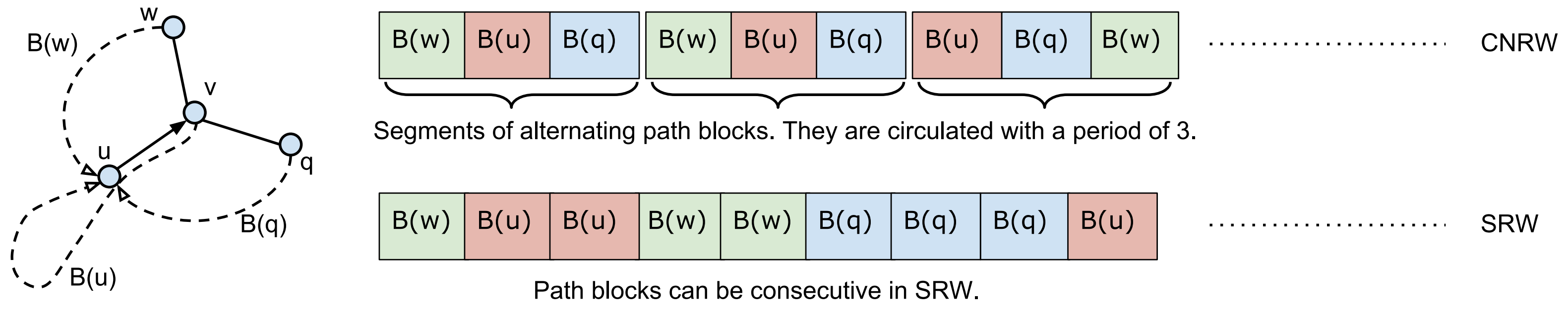

B(w)
w
v
u
q
B(q)
B(u)
B(w) B(u) B(q) B(w) B(u) B(q) B(u) B(q) B(w)
CNRW
Segments of alternating path blocks. They are circulated with a period of 3.
B(w) B(u) B(u) B(w) B(w) B(q) B(q) B(q) B(u)
SRW
Path blocks can be consecutive in SRW.

Figure : Comparisons of the block distribution in CNRW and SRW. CNRW creates alternating stratified path blocks that boost the sampling performance.

GNRW: Groupby Neighbors Random Walk

- **Key idea.** GNRW is a natural extension of CNRW. Instead of performing the circulation at the granularity of each neighbor (of v), we propose to first *stratify* the neighbors of v into groups, and then circulate the selection among all groups.
- **Theorems**
  - Theorem 1. GNRW has the same stationary distribution $\pi(v) = k_v/2|E|$ as SRW's.
  - Theorem 2. The asymptotic variance of GNRW is no greater than SRW's.

$$V_\infty(\hat{\mu*}) \leq V_\infty(\hat{\mu}).$$

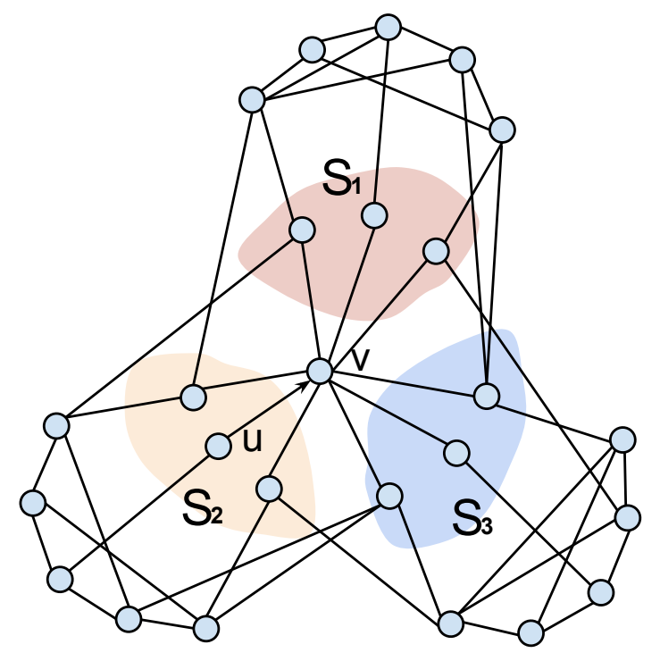


Figure : An example of partitioning a node's neighbors into 3 groups. GNRW chooses a next group by circulating $\{S_1, S_2, S_3\}$.

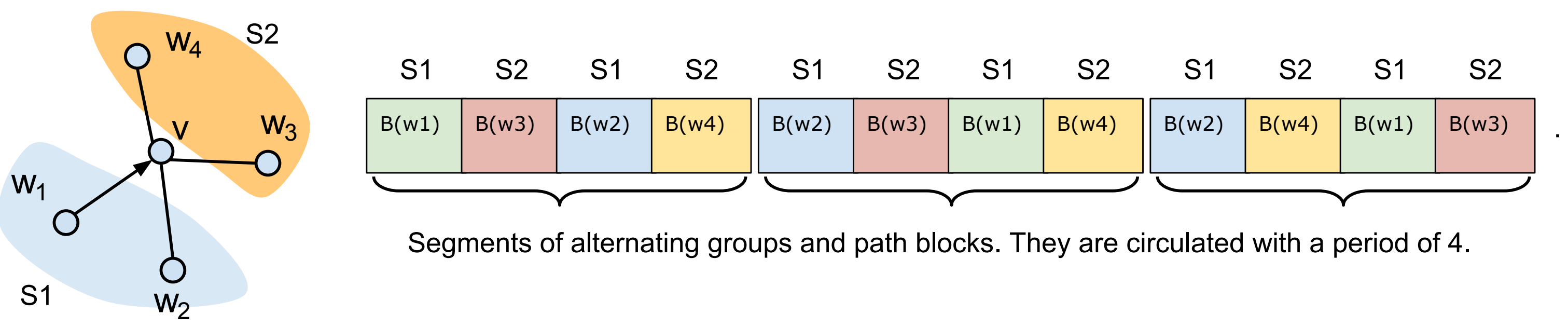
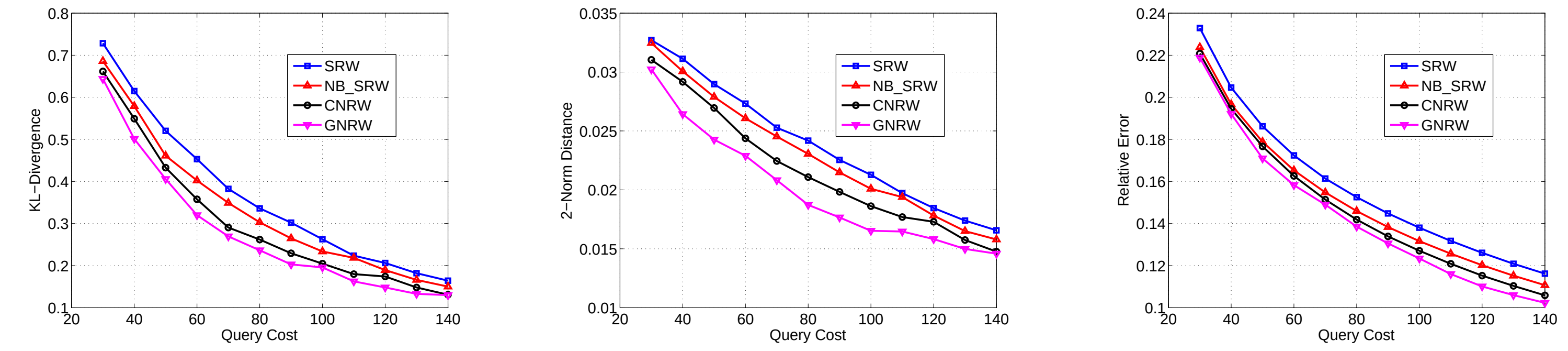

W4
S2
V
W3
W1
S1
W2
S1 S2 S1 S2 S1 S2 S1 S2 S1 S2 S1 S2
B(w1) B(w3) B(w2) B(w4) B(w2) B(w3) B(w1) B(w4) B(w2) B(w4) B(w1) B(w3) .....
Segments of alternating groups and path blocks. They are circulated with a period of 4.

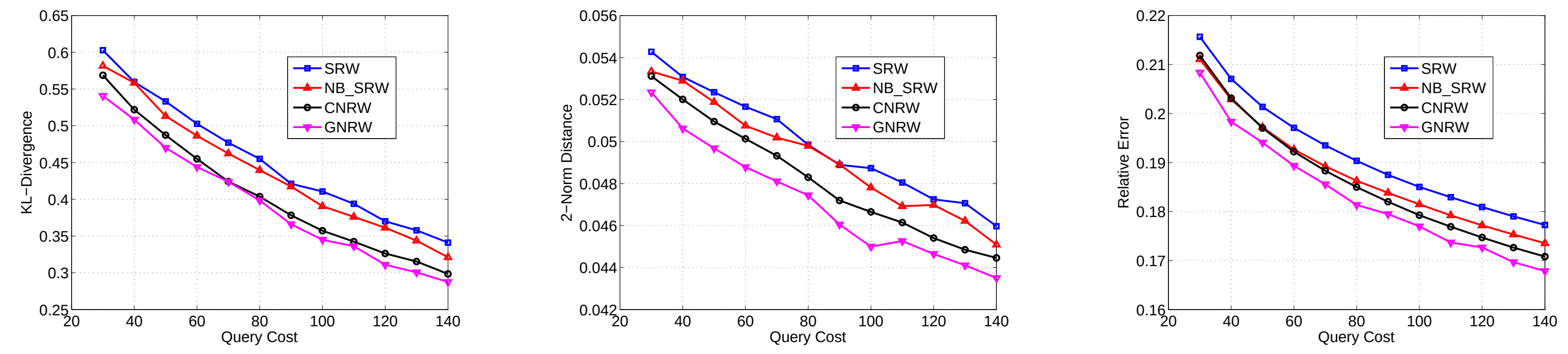
Figure : A demo of GNRW: it makes higher order of stratifications of the path blocks by grouping them into strata.

Experimental Results



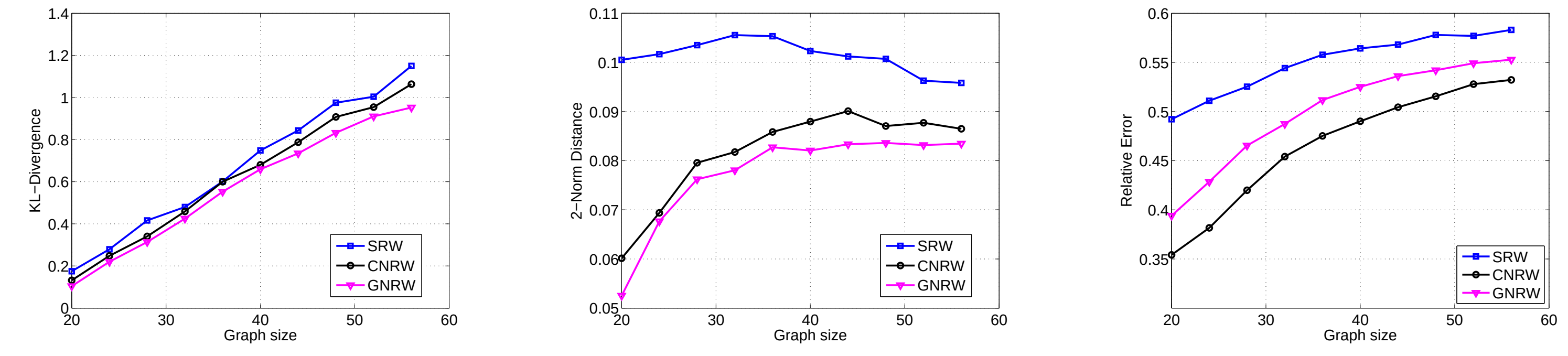
(a) KL-divergence (b) ℓ_2-distance (c) Estimation error

Figure : Public benchmark dataset: Facebook



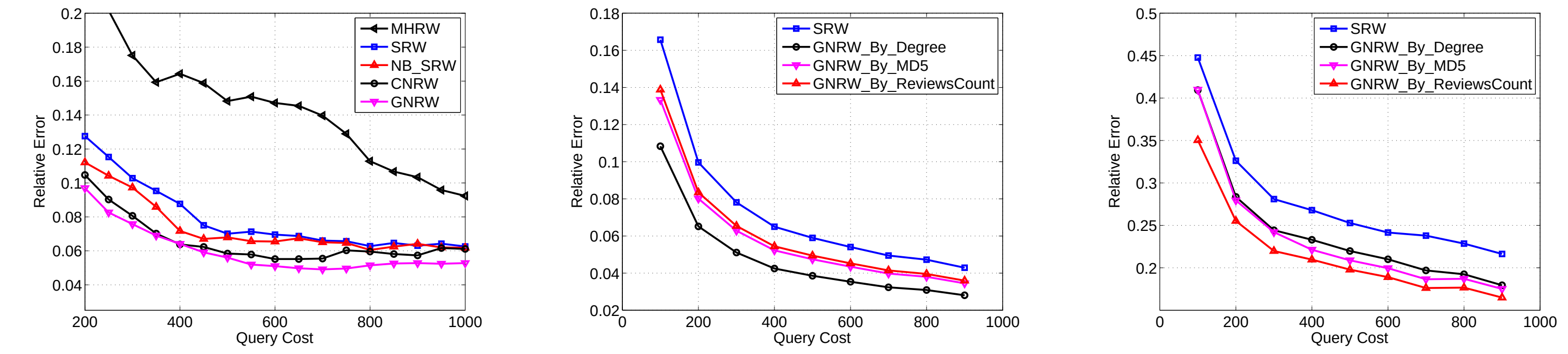
(a) KL-divergence (b) ℓ_2-distance (c) Estimation error

Figure : Synthetic datasets: clustered graph



(a) KL-divergence (b) ℓ_2-distance (c) Estimation error

Figure : Barbell graph size analytics.



(a) Google Plus (b) Yelp (average degree) (c) Yelp (average reviews count)

Figure : Performance comparisons on large OSNs.

Conclusion

In this paper, we developed two algorithms: (1) CNRW, which replaces the memoryless transition in simple random walk with a memory-based, sampling-without-replacement, transition design, and (2) GNRW, which further considers the observed attribute values of neighboring nodes in the transition design. We proved that while CNRW and GNRW achieve the exact same target (sampling) distribution as traditional simple random walks, they offer provably better (or equal) efficiency no matter what the underlying graph topology is.

Selected References

[1] E. M. Airoldi. Sampling algorithms for pure network topologies. SIGKDD Explorations, 7:13-22, 2005.

[2] N. Alon. Eigenvalues and expanders. Combinatorica, 6:83-96, 1986. 10.1007/BF02579166.

[3] S. Boyd, P. Diaconis, and L. Xiao. Fastest mixing markov chain on a graph. SIAM REVIEW, 46:667-689, 2003.

[4] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In INFOCOM, 2010.

[5] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou. Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In SIGMETRICS, 2011.